WATERMARK ESTIMATION THROUGH DETECTOR ANALYSIS

Ton Kalker *Jean-Paul Linnartz* *Marten van Dijk*

Philips Research
Prof. Holstlaan 4
5656 AA Eindhoven
The Netherlands
{kalker,linnartz,mvandijk}@natlab.research.philips.com

ABSTRACT

A watermark is a perceptually unobtrusive signal embedded in an image, an audio or video clip, or any other other multimedia asset. Its purpose is to be a label which is holographically attached to the content. Moreover, it can only be removed by malicious and deliberate attacks (without a great loss of content quality) if some secret parameter K is known. In contrast, a watermark should be readily detectable by electronic means. This implies that electronic watermark detection is only feasible if the watermark detector is aware of the secret K. In many watermarking business scenarios the watermark detector will be available to the public as a black box **D**. *The following question is therefore justified: can the secret K be deduced from the operation of the black box* **D**? *And if yes, what is the complexity of this process? In this paper we will address these questions for a large class of watermarking schemes. This work is an extension of earlier work at Philips Research [1].*

1. INTRODUCTION

Watermarking is a fundamental enabling technology for the distribution of digital multimedia (MM) content. At present it is very easy to distribute and copy digital multimedia content. Without any special precautions the content generation and distribution industry will be very reluctant to publish in the digital domain. The slow introduction of the new Digital Versatile Disk (DVD) format bears witness to this tendency.

Digital watermarking is a technical solution to the copyright problem. In its basic form a digital watermark W is a *small* signal added to MM content. The watermark W carries sufficient data to ensure proper copyright verification. Due to its intended purpose a watermark should be unobtrusive (i.e. no perceptible degradation of the quality is allowed), easily detectable by dedicated software or hardware and very difficult to remove by malicious and deliberate attacks.

It is essential to distinguish two types of applications of watermarking technology. In the first type of application all content can be enforced to contain a watermark. A typical example is given by (images on) bank notes and smart cards. It is not sufficient for a pirate to remove the watermark (i.e. reconstruct the original content), but she will actually have to insert a watermark which contains false copyright information. By relying on cryptographic methods the complexity of this type of attack can be made arbitrarily large.

In the second type of application watermarking cannot be enforced. A typical example is given by film content on DVD. The film industry can enforce watermarks on commercial digital video, but it cannot enforce watermarking of home videos. Therefore DVD players will have to accept both watermarked (i.e. copyright protected) and unwatermarked content. This implies that it is sufficient for a pirate to remove a watermark from a commercial video (i.e. make a good estimate of the unwatermarked original) in order to invalidate the copyright protection mechanism of DVD. In this paper we will focus on this type of non-watermark-enforced application.

An obvious requirement of any watermarking scheme is that it is not feasible for a pirate to reconstruct a good approximation of the original content having only one watermarked copy at her disposal (a *one-copy* attack). This is a relevant issue as shown by the case of cartoons or graphics content. This type of content is characterized by sparse histograms. Any watermarking scheme which affects this sparseness is vulnerable to *one-copy* attacks.

A business scenario which allows several (differently) watermarked copies of one original on the market is also vulnerable to attacks. By simply averaging over a large set of copies a good unwatermarked ap-

0-8186-8821-1/98 $10.00 © 1998 IEEE

proximation of the original may be obtained [2] [9]. In this paper we will study a third type of security risk, viz. the availability to a pirate of a watermark detector. A typical example is given by DVD where a copyright protection system based upon watermarking will imply a watermark detector in every single DVD player. To our knowledge this problem has first been studied in [4] and [1].

The essence of any watermarking scheme is the modification of signal values. These changes are effectuated in the spatial domain [5], the DCT domain [6], the wavelet domain [7], the FFT domain or more exotic domains [8] [3]. In many cases watermark detection amounts to thresholding a correlation computation. A suspect signal I_s (or an extracted set of features) is correlated with some pattern W to obtain a correlation value $d = \langle I_s, W \rangle$. If this value d is larger than a signal dependent threshold $\tau(I_s)$ then the watermark is said to be present. Otherwise the watermark is said to be absent. It can easily be shown that knowledge of the pattern W allows a pirate to annihilate a watermark from an arbitrary signal.

In this paper we study a simplified version of this detection process from the pirates point of view. We assume that the pirate has a general knowledge of the operation of the watermark detector, but not of the secret pattern W. Motivated by the discussion above we pose the following questions:

1. How much knowledge can a pirate obtain about the patterns W by observing the behavior of the detector? Here we assume that she can offer any signal of her own choosing to the detector.

2. What is the complexity of the above process? That is, how many experiments are needed to derive sufficient knowledge of W?

This paper is organized as follows. In Section 2 we present a simplified mathematical model of the watermark detection process. This allows us to formulate the two questions above in a mathematical precise manner. In Section 3 we present a method for retrieving the secret pattern W. In Section 4 this method is experimentally validated. Section 5 summarizes the paper and gives recommendations for improving watermark security.

2. MATHEMATICAL FORMULATION

As introduced above we can view a watermark detector box **D** as a quantized correlator. In this section we will formulate a simplified model of such a watermark detector, but with slight modifications many proposed watermark detectors will fit this description.

In our simplified model the detector **D** is characterized by a number of parameters: the length N of the input vectors, two decision thresholds A and B, and a *DC-free* random vector $w = \{w_i\}_{i=0}^{N-1}$, $w_i \in \{-1, +1\}$. As usual DC-freeness is defined by $\sum_i w_i = 0$. The detector **D** takes as input a vector $x = \{x_i\}_{i=0}^{N-1}$ and gives as output the symbol "1" (signal x contains the watermark w) or "-1" (signal x does not contain the watermark w).

The detector **D** implements a *deterministic* algorithm to come to its decision. It starts with normalizing the signal x, followed by the computation of a normalized correlation value d. If this value d is large enough, i.e. larger than B, the value 1 is returned. If the value d is too small, i.e. smaller than A, the value -1 is returned. If the value of d is in between A and B, a value -1 or 1 is returned depending on the value of some *hash* function hash(x) applied to the signal x. The probability that a value 1 is returned then depends monotonically on the value of d. If $d = A$ this probability equals 0, if $d = B$ this probability equals 1.

Prime examples of watermarking which use slight variations of this kind of detection scheme are proposed in [10] and [5]. These schemes have $A = B$ and use a somewhat different (but still discrete) value set for the watermark samples w_i. The theory can easily be extended to cover more advanced proposals [11] [6] but for reasons of simplicity of presentation we stick to the above model.

The only way an attacker can obtain information from a watermark detector **D** is by observing changes in detection behavior when offering different signals as input. The bigger the changes observed the more information can be obtained. For this reason a good watermark detector will not use a single threshold $A = B$, but it will use an uncertainty interval $[A, B]$. In fact, it is argued in [1] that from the standpoint of watermark vulnerability the probability function on the interval $[A, B]$ should be equal to a squared sine function. Therefore the detector **D** implements a $\{-1, 1\}$-valued *randomizing* function $f(y, h)$ such that for a fixed value of $0 \leq y_0 \leq 1$ the probability that $f(y_0, h) = 1$ is equal to $\sin^2(\frac{\pi x_0}{2})$ (when the hash value h ranges over all allowed values).

In pseudo-code the algorithm implemented by **D**

reads as follows:

```
x := x − mean_value(x);
x := x/ standard_deviation(x);
d := (∑_{i=0}^{N−1} x_i w_i)/√N ;
if d < A
    return −1;
elseif d > B
    return 1;
else
    return f((d − A)/(B − A), hash(x));
fi
```

In order to analyze the security of this detection scheme we will assume that an attacker has a detector box **D** and a watermarked signal x_0 available. The attacker has no knowledge of the vector w, the thresholds A and B, the hash function "hash" and the randomizing function f (other than the characterizing probability property). Referring to Section 1, we again pose the questions of how much knowledge she can obtain about the vector w and what effort is needed to obtain this information.

3. ATTACK

We propose the following attack on the detection scheme as described in Section 2.

1. Construct a signal x_1 such that its normalized correlation value lies halfway between A and B. The signal x_1 can be obtained from x_0 by iteratively replacing sample values of x_0 by the mean value of x_0.

2. Initialize a set of counters $c_i = 0, i = 0, \ldots, N-1$.

3. Choose a random *DC-free* vector $v = \{v_i\}$, $v_i \in \{-1, +1\}$.

4. Form the signal $x_1 + s * v$ and record the decision $d \in \{-1, 1\}$ by the detector. The *strength* parameter s is determined experimentally by observing the growth of the counters c_i. If s is chosen too small or too large the distribution of counter values will be more or less normally distributed with mean 0. For s chosen properly, the distribution of counter values will be a sum of two normal distributions with equal variance but with mean values of opposite sign.

5. Update all counters by the rule

$$c_i := c_i + d * v_i.$$

6. Go back to 3 until all counters are sufficiently different from 0 or until the loop has been traversed a predetermined number of times.

7. Estimate w by $w_i = \text{sign}(c_i)$.

In [4] and [1] a different version of this attack has been presented. The overall conclusion of both approaches however is the same: watermark detection schemes which are modeled as above can be cracked in $\mathcal{O}(N)$ experiments (i.e loop traversals).

4. EXPERIMENTS

The attack described in Section 3 has experimentally been verified for a watermark detector for images. An attack was simulated in MATLAB for a watermarked image[1] of size 128×128 (marked at 9 times standard deviation) and a software model of a watermark detector with $A = 5$ and $B = 7$. First an image at the threshold of detection was obtained by iteratively replacing sample values by the mean value of the image (see Figure 1). Then for each s in a set of strength parameters $s = 1, 3, 5, 7, 9, 11, 13, 15$ the attack loop as described in Section 3 was traversed $5 \times 128 \times 128$ times. The results of this experiment are given in Table 1. This table clearly shows that for a proper choice of the strength parameter s, the detector secret w can be retrieved with great accuracy for a relatively small effort. Note that a random estimation of the watermark will lead to a 50% retrieval percentage and that a perfect estimation will lead to 100% retrieval percentage.

In a more extensive experiment both the strength parameter s and the number of iterations were simultaneously varied. Figure 2 shows the percentage of correctly estimated watermark values as a function of these two parameters. The figure clearly shows that increasing the number of iterations always improves the reliability of retrieval, although with diminishing returns. The figure also clearly shows that increasing the perturbation strength has its limits, and that beyond a certain value of s the retrieval percentage drops. This can easily be understood by noting that for large perturbations the original (watermarked) signal is lost in the overall noise.

In other experiments the threshold s was fixed at 10 but B was varied from 5 to 7.4 in steps of 0.3. The results are summarized in Table 2. This experiment confirms the prediction of both [1] that more effort is needed for larger uncertainty intervals $[A, B]$ (without affecting the property of linear complexity though).

[1]The central part of the well known Lena image was chosen for this experiment.

Figure 1: The (partial) Lena image at detection threshold.

strength	1	3	5	7
percentage	53.8%	60.5%	66.5%	72.3%
strength	9	11	13	15
percentage	75.8%	78.4%	80.0%	80.1%

Table 1: Percentage of watermark retrieval as function of the strength parameter s.

5. CONCLUSIONS

In this paper we have addressed the issue of watermark security based on the availability of a watermark detector and a single watermarked signal. We extended the work started in [4] and [1] by presenting a more simple attack method. We have shown that for DC-free and $\{-1, +1\}$-valued watermarks the complexity of watermark retrieval is linear in the number of sample values. A slight extension of the theory shows that the same method is applicable to all correlation based watermark methods. This implies that watermarking methods based upon thresholded correlation are not suited for applications where watermarking cannot be enforced (DVD).

interval width	0.3	0.6	0.9	1.2
percentage	95.6%	94.0%	91.5%	88.0%
interval width	1.5	1.8	2.1	2.4
percentage	82.7%	79.1%	76.4%	72.4%

Table 2: Percentage of watermark retrieval as function of the width of the uncertainty interval $[A, B]$

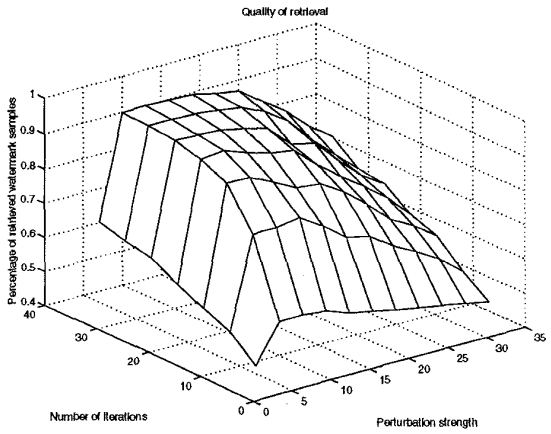

Figure 2: Percentage of retrieval as a function of the strength parameter s and the number of iterations.

It can be shown that the core problem with the attack described above lies with the pseudo-linear operation of the watermark detection scheme. It follows that any watermarking scheme suitable for non-watermark-enforced applications needs strongly non-linear ingredients.

6. REFERENCES

[1] J.P. Linnartz and M. van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In *Pre-Proceedings of the Workshop on Information Hiding*, Portland, April 1998.

[2] H. Stone. Analysis of attacks on image watermarks with randomized coefficients. Technical report, NEC Research Institute, May 1996.

[3] M.J.J. Maes and C.W.A.M. van Overveld. Digital watermarking by geometric warping. In *Proceedings of the ICIP*, Chicago, October 1998.

[4] I.J. Cox and J.P.M.G. Linnartz. Public watermarks and resistance to tampering. In *Proceedings of the ICIP*, Santa Barbara, California, October 1997. Paper appears only in CD version of proceedings.

[5] I. Pitas. A method for signature casting on digital images. In *Proceedings of the ICIP*, volume 3, pages 215 – 218, Lausanne, September 1996.

[6] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for images,

audio and video. In *Proceedings of the ICIP*, volume 3, pages 243 – 246, Lausanne, Switzerland, September 1996.

[7] X.-G. Xia, C.G. Boncelet, and G.R. Arce. A multiresolution watermark for digital images. In *Proceedings of the ICIP*, volume 1, pages 548 – 551, Santa barbara, California, October 1997.

[8] J.J.K. Ó Ruanaidh and T. Pun. Rotation, scale and translation invariant digital image watermarking. In *Proceedings of the ICIP*, volume 1, pages 536 – 539, Santa barbara, California, October 1997.

[9] M.J.J. Maes. Twin peaks: The histogram attack to fixed depth image watermarks. In *Pre-Proceedings of the Workshop on Information Hiding*, Portland, April 1998.

[10] W. Bender, D. Gruhl, and N. Morimoto. Techniques for data hiding. In *Proceedings of the SPIE*, volume 2420, page 40, San Jose CA, USA, February 1995.

[11] G.W. Braudaway. Protecting publicly-available images with an invisible image watermark. In *Proceedings of the ICIP*, volume 1, pages 524 – 527, Santa Barbara, California, October 1997.